Reputation Equity in Ranking Systems

Guilherme Ramos
Dept. of Electrical and Computer Engineering - University of Porto
Porto, Portugal
guilhermeramos21@gmail.com

Ludovico Boratto
University of Cagliari
Cagliari, Italy
ludovico.boratto@acm.org

Mirko Marras
EPFL
Lausanne, Switzerland
mirko.marras@acm.org

ABSTRACT

The impact of ranking systems on humans is an aspect that is getting a lot of attention. In this paper, we consider a class of algorithms, known as *reputation-based ranking systems*, which rank the items based on a reputation score automatically computed for each user. Recent literature introduced the concept of *reputation independence*, which considers a sensitive attribute of the users (such as gender or age) and makes the reputation scores independent from that attribute. Here, we show that if we consider a different sensitive attribute w.r.t. a user to introduce independence, reputation scores are still biased. To overcome this issue, we propose an approach to attain equity in the reputation scores computation, independently of *any* sensitive attribute that characterizes the users.

CCS CONCEPTS

• **Information systems** → **Learning to rank**; • **Applied computing** → *Law, social and behavioral sciences.*

KEYWORDS

Ranking Systems, Reputation Systems, Bias Mitigation.

ACM Reference Format:
Guilherme Ramos, Ludovico Boratto, and Mirko Marras. 2021. Reputation Equity in Ranking Systems. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482171

1 INTRODUCTION

Accounting for the impact of ranking systems on humans has become a topic of prime relevance in information retrieval. Users consider highly ranked results as more relevant [6]. So, any bias in the ranking can lead to negative consequences, such as users' discrimination [1–3, 10, 12, 14–16], polarization [6], or bribing [9, 11].

Humans can be affected by a ranking, not only when they are the end-users exploring the results (e.g., in a Web search), or when they are the ones being ranked (e.g., as job candidates). Indeed, a class of algorithms, known as *reputation-based ranking systems*, scores the users to decide how to rank items [7, 11]. Recent work by Ramos and Boratto [8] has shown that reputation scores are biased on users' sensitive attributes, leading to minority demographic groups (such as females) being deemed as less relevant, thus accounting for their preferences less when computing items' rankings (*disparate reputation*). To overcome disparities, the concept of *reputation independence* (RI) was introduced, ensuring that the reputation scores of users belonging to different legally-protected groups are statistically indistinguishable. However, the work in [8] makes interventions on individual sensitive attributes and, as highlighted by Kleinberg et al. [5], this does not provide guarantees to groups obtained considering multiple sensitive attributes.

In this paper, we study if RI on a single sensitive attribute (e.g., gender) provides RI to groups shaped considering a different sensitive attribute (e.g., age). We show, theoretically and experimentally, that introducing independence on one attribute still leads to disparate reputation when considering a different one. For this reason, we propose an approach to achieve equity in the way reputation scores are computed [13] by making them independent of the demographic attributes that characterize the users. With our approach, each user has a different reputation score (thus weighing their preferences differently), but (multi-attribute) demographic groups will have a similar reputation when shaping the ranking. Hence, we ensure that each group contributes equally to a ranking. Experiments on real-world data show that our approach can remove bias from multiple attributes in the reputation scores.

Our contributions are as follows: (i) we provide evidence, both theoretically and experimentally, that, for RI to be guaranteed, it should cover multiple sensitive attributes of the users; (ii) we extend the existing notion of RI, to cover more than one sensitive attribute, and study its complexity; and (iii) we evaluate the capability of our approach at creating rankings based on less biased reputations and compare it against state-of-the-art reputation debiasing solutions.

2 PRELIMINARIES

Context formalization. Given a set $\mathcal{U} = \{u_1, \ldots, u_n\}$ of $n \in \mathbb{N}$ users and a set $\mathcal{I} = \{i_1, \ldots, i_m\}$ of $m \in \mathbb{N}$ items, a user $u \in \mathcal{U}$ can rate an item $i \in \mathcal{I}$. This feedback shapes a rating matrix, $\mathcal{R} \in \mathbb{R}^{n \times m}$. Ratings are normalized to be in $]0, 1]$. The difference between the maximum and the minimum normalized ratings is denoted by Δ_R. We denote the ranking of an item i by r_i, with $r_i \in]0, 1]$, that corresponds to a score of the item computed from the item ratings.

We consider $\mathcal{A} = \{A_1, \ldots, A_K\}$ as a set of $K > 0$ user attributes (e.g., gender, age) and let each attribute $A_j = \{a_{j_1}, \ldots, a_{j_{s_j}}\}$, with $1 \leq j \leq k$, have $s_j \in \mathbb{N}$ classes. For instance, an attribute A_j abstracting user's genders can have two or more classes, i.e., $A_j = \{male, female, \ldots\}$. We denote classes of an attribute $A_j \in \mathcal{A}$ by a_j, a'_j or, when we want to enumerate them, by $a_{j_1}, \ldots, a_{j_{s_j}}$, where s_j is the number of classes of attribute A_j, and we assume that $A_j(u) = a_j$ is the class $a_j \in A_j$ for attribute $A_j \in \mathcal{A}$ a user $u \in \mathcal{U}$

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482171

belongs to. We denote the set of users who rated item $i \in \mathcal{I}$ by $\mathcal{U}_i = \{u \in \mathcal{U} : R_{ui} > 0\}$, the set of items that user $u \in \mathcal{U}$ rated by $\mathcal{I}_u = \{i \in \mathcal{I} : R_{ui} > 0\}$, and the set of users in class $a_j \in A_j$ of attribute $A_j \in \mathcal{A}$ by $\mathcal{U}(a_j) = \{u \in \mathcal{U} : A_j(u) = a_j\}$. If an attribute $A_j \in \mathcal{A}$ has classes $A_j = \{a_{j_1}, \dots, a_{j_{s_j}}\}$, we assume that $\mathcal{U}(a_j) \cap \mathcal{U}(a'_j) = \emptyset$ for all $a_j, a'_j \in A_j$, with $a_j \neq a'_j$. For a vector $v \in \mathbb{R}^n$, $avg(v)$ denotes its *average* and $std(v)$ its *standard deviation*. *Reputation-based ranking.* We focus on a class of ranking systems that assign a relevance score to a user (*reputation*). These systems rank items by weighing user preferences with each user's reputation. Li et al. [7] proposed a reputation-based system implementing an iterative method with exponential rate convergence. Saúde et al. [11] extended this scheme. At each iteration, their scheme updates the ranking of each item i, r_i^{k+1}, as a weighted average of ratings given to i with the reputations, c_u^k, of users that rated the item; the system updates the users' reputation by computing how much user's ratings disagree to items' ranking. Formally:

$$\begin{aligned} r_i^{k+1} &= \sum_{u \in \mathcal{U}} R_{ui} c_u^k \Big/ \sum_{u \in \mathcal{U}} c_u^k \\ c_u^{k+1} &= 1 - \frac{\lambda}{|\mathcal{I}_u|} \sum_{i \in \mathcal{I}_u} |R_{ui} - r_i^{k+1}| \end{aligned} \qquad (1)$$

for any initial $c_u^0 \in]0,1]$ (we select $c_u^0 = 1$) and for $\lambda \in]0,1]$, a hyper-parameter that penalizes the discordance of a user given ratings with the items' rankings.

Problem formalization. Given a set of users $\mathcal{U}$, a set of items $\mathcal{I}$, a set of ratings given by users to items $\mathcal{R}$, and a set of user's attributes $\mathcal{A}$ such that $A_j = \{a_{j_1}, \dots, a_{j_{s_j}}\} \in \mathcal{A}$, our goal is to: **A** – compute users' reputation $\{c_u\}_{u \in \mathcal{U}}$ on user preferences, capturing how relevant are the preferences of a user for the community as a whole, in a ranking system; **B** – compute rankings of items $\{r_i\}_{i \in \mathcal{I}}$ as a weighted average of the users' reputations and the items' ratings; and **C** – obtain reputations' distributions that, for every K-tuple pair of classes $(a_1, \dots, a_j) \in A_1 \times \ldots \times A_K$, each associated to the set of users $\mathcal{U}(l = (a_1, \dots, a_K))$, are statistically indistinguishable.

3 SINGLE-ATTRIBUTE RI

3.1 Single-Attribute Mitigation Methodology

Ramos and Boratto [8] introduced the *Disparate Reputation* (DR) concept in reputation-based ranking systems, as the difference between the average reputation of the users characterized by two distinct classes, a and b, for the same attribute: $\Delta(a,b) = \mu_a - \mu_b$. The metric ranges in $[-1+\Delta_R\lambda, 1-\Delta_R\lambda]$. Its value is 0 when $\mu_a = \mu_b$.

To characterize if DR systematically affects users in a class, the authors performed a Mann-Whitney (MW) statistical test. The test was performed between each pair of user groups' reputation distributions, relative to an attribute. To mitigate DR when considering a single users' sensitive attribute (that does not need to be binary), the authors performed the following final post-processing step.

$$\begin{aligned} c_u &= \mu + \left(c_u^N - \mu_l\right) \frac{\sigma}{\sigma_l}, \text{for } l = 1, \dots, K \text{ and } u \in \mathcal{U}(a_l) \\ r_i &= \sum_{u \in \mathcal{U}} R_{ui} c_u \Big/ \sum_{u \in \mathcal{U}} c_u \end{aligned} \quad , \qquad (2)$$

where μ_l and σ_l are the average and standard deviation of users' reputation for users $u \in \mathcal{U}(a_l)$.

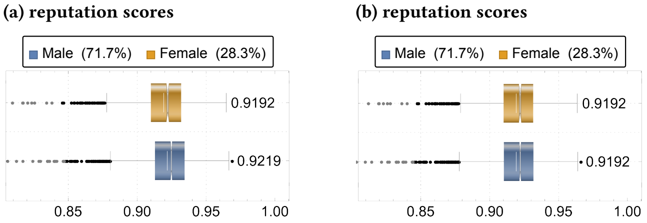

Fig. 1: [ML-1M] Box-whisker-chart for users' reputations after (1), (a), and after (1) and (2), (b), with $\lambda = 0.5$, for gender-based groups.

The authors showed that with this step the reputations' distributions for each class of a sensitive attribute become statistically indistinguishable, leading to single-attribute RI. However, it is unclear whether an analogous bias on other attributes ends up being mitigated when only one attribute is considered. This motivated us to make a more extensive evaluation of the described methodology.

3.2 Evaluation on Multiple Sensitive Attributes

We start by doing an exploratory analysis of the reputation-based (RepRank) system formalized in Section 2. The goal is to understand if, when mitigating bias for a sensitive attribute, there is still a bias related to another sensitive attribute. We consider the ML-1M dataset [4], which includes gender and age as user's sensitive attributes. More details on this dataset can be found in Section 5.1.

First, we apply the mitigation described in (2), grouping users on their gender. The results on DR (Fig. 1) confirm that introducing the extra step in (2) in the RepRank method leads to users' reputations independence for gender-based groups. Notwithstanding, if we group reputation scores based on another sensitive attribute – the age – and measure the DR on the resulting reputation distributions, then there is a DR for the attribute age, as shown in Fig. 2. By applying (1) and (2) sequentially, we mitigate a reputation bias on the attribute gender. However, as depicted in Fig. 2, the reputation bias on age groups is not mitigated in a collateral fashion. The sub-Figs. of Fig. 2 indicate the reputations on age-based groups before and after mitigating for attribute gender, showing almost identical values. Hence, mitigating for attribute gender does not mitigate for attribute age collaterally. We also performed analyses (not reported for space constraints) with (2) to mitigate reputation bias on age and test gender reputation bias, still observing disparities.

Proposition 1. *Given a set of users $\mathcal{U}$, a set of items $\mathcal{I}$, a set of ratings $\mathcal{R}$, and a set of user attributes $\mathcal{A} = \{A_1, \dots, A_K\}$, mitigating a reputation bias with (2) for each attribute (for any order) does not necessarily yield reputations without bias for both attributes.* ○

Proof. We give a counter-example. For the ML-1M dataset and attributes gender and age, we apply (2) to mitigate reputation bias for gender first, and after for age. Gender reputations' averages are statistically different, $\mu_{female} = 0.906088 \neq \mu_{male} = 0.906067$. □

4 MULTI-ATTRIBUTE RI

We design a strategy that, given a set of users' sensitive attributes, mitigates the user reputations' bias against user groups characterized by different combinations of those attributes.

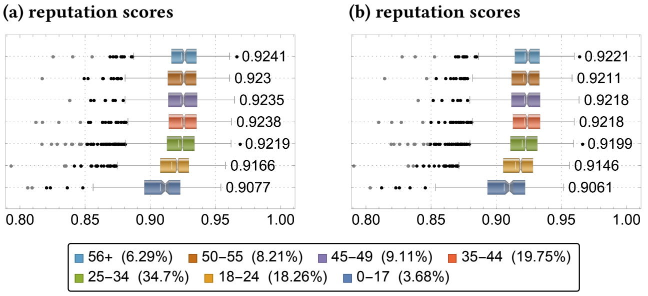

Fig. 2: [ML-1M] Box-whisker-chart for reputations of users resulting from (1) (a), and from (1) and (2) applied to attribute gender and evaluated on attribute age (b), with $\lambda = 0.5$, for groups based on age.

The method partitions users according to more than one attribute, jointly. Specifically, let $\mathcal{A} = \{A_1, \dots, A_K\}$ be a set of $k > 0$ attributes and let each attribute $A_j = \{a_{j,1}, \dots, a_{j,s_j}\}$ have s_j classes. Now, we consider all the K-tuples of classes $(a_1, \dots, a_j) \in A_1 \times \dots \times A_K$. Subsequently, to each K-tuple, we associate the set of users $\mathcal{U}(l = (a_1, \dots, a_K))$, which is the set of users such that $a_j \in A_j$ for $j = 1, \dots, K$. Observe that the sets of users for all the possible K-tuples form a partition of U, as desired. Hence, we arrange (2) as

$$\begin{aligned} c_u &= \mu + \left(c_u^N - \mu_l\right)\frac{\sigma}{\sigma_l}, && \text{for } u \in U(l = (a_1, \dots, a_K)) \text{ and} \\ & && (a_1, \dots, a_K) \in A_1 \times \dots \times A_K \\ r_i &= \sum_{u \in \mathcal{U}} R_{ui} c_u \Big/ \sum_{u \in \mathcal{U}} c_u \end{aligned} \tag{3}$$

where, for $l \in A_1 \times \dots \times A_K$, $\mu = \min_l \mu_l$ and $\sigma = \min_l \sigma_l$, with $\mu_l = avg\left(\{c_u^N\}_{u \in U(l)}\right)$ and $\sigma_l = std\left(\{c_u^N\}_{u \in U(l)}\right)$. Note that, in (3), choosing the minimum of the averages and the minimum of the standard deviations ensures that reputations' re-scaling lies in the interval $]0, 1]$. Similarly to (2), even (3) reconciles reputation's distributions for each K-tuple of attributes classes so that the reputations of each K-tuple of classes are "statistically indistinguishable". This leads to the targeted *multi-attribute RI*.

Remark 1. *The post-processing in* (3) *can be used in any ranking system computing rankings as a weighted average of the ratings.* ⋄

Theorem 1. *Consider a matrix of ratings $\mathcal{R}$, with set of items $\mathcal{I}$, set of users $\mathcal{U}$ and set of K users' sensitive attributes $\mathcal{A} = \{A_1, \dots, A_K\}$. Let the users' reputations and items' rankings be computed with* (1). *If we apply* (3) *to recompute users' reputations and items' rankings, using the attributes $A_1, \dots, A_K$, then the following property holds: for any two classes of any two sensitive attributes, $a \in A_i$ and $a' \in A_j$ ($A_i, A_j \in \mathcal{A}$), the set of users $\mathcal{U}(a)$ reputations and the set of users $\mathcal{U}(a')$ reputations have zero DR ($\mu_a = \mu'_a$).* ○

Proof. First, for any two classes of any two sensitive meta-attributes $\tilde{a} \in A_1 \times \dots \times A_K$ and $\tilde{a}' \in A_1 \times \dots \times A_K$, (3) makes the set of users $\mathcal{U}(\tilde{a})$ reputations and the set of users $\mathcal{U}(\tilde{a}')$ reputations have zero DR ($\mu_{\tilde{a}} = \mu_{\tilde{a}'} = \mu$). For each set of users $\mathcal{U}(a_1, \dots, a_K)$, as noted before, the average reputation is the same, μ. It remains to show that the average of a finite collection of finite sets with the same average μ is also μ. To ease the notation, consider the sets of users $U_1, \dots, U_k$ that have the same average reputation μ. We observe that: $\frac{\sum_{i=1}^k \sum_{u \in U_i} c_u}{\sum_{i=1}^k |U_i|} = \frac{\sum_{i=1}^k |U_i| \frac{\sum_{u \in U_i} c_u}{|U_i|}}{\sum_{i=1}^k |U_i|} = \mu \frac{\sum_{i=1}^k |U_i|}{\sum_{i=1}^k |U_i|} = \mu.$ □

Proposition 2. *Given a set of users $\mathcal{U}$, a set of items $\mathcal{I}$, a set of ratings that users gave to items $\mathcal{R}$ and a set of user attributes $\mathcal{A} = \{A_1, \dots, A_K\}$, the time-complexity of computing the iterative scheme in* (1) *for $N > 0$ iterations followed by* (3) *is $O\left(N|\mathcal{U}||\mathcal{I}| + k|\mathcal{U}|\right)$.* ○

Proof. First, we perform (1) for N iterations, with each iteration composed of two steps. The first step computes the rankings of $|\mathcal{I}|$ items as a weighted average of the users' ratings by the users' reputations, which takes $O(|\mathcal{U}|)$ time. For one iteration, the rankings' update requires $O(|\mathcal{I}||\mathcal{U}|)$. The second step of each iteration requires to compute $|\mathcal{U}|$ users reputations. Updating one user u reputation implies running $\mathcal{I}_u \subseteq \mathcal{I}$ operations. The second step has $O(|\mathcal{I}||\mathcal{U}|)$ operations as well. For N iterations, the time-complexity is $O(N|\mathcal{I}||\mathcal{U}|)$. We then compute (3) in two steps. The first step of (3) requires updating $|\mathcal{U}|$ users' reputations. We can compute μ_l and σ_l linearly in $|\mathcal{U}(l)|$. Since $\bigcup_l \mathcal{U}(l) = \mathcal{U}$ is a users' partition, we can compute all the μ_l and σ_l in $O(|\mathcal{U}|)$. Also, we may verify if $u \in \mathcal{U}(l = (a_1, \dots, a_K))$ in $O(K)$, by checking the K classes values of the K attributes in a table indexed by the users. The second step of (3) updates the items' rankings in $O(|\mathcal{I}||\mathcal{U}|)$ time. Hence, (3) requires $O(K|\mathcal{U}| + |\mathcal{I}||\mathcal{U}|)$ steps. In total, $O\left(N|\mathcal{U}||\mathcal{I}| + K|\mathcal{U}|\right)$. □

In general, $K \leq |\mathcal{I}|$, and the time-complexity of Proposition 2 becomes $O\left(N|\mathcal{U}||\mathcal{I}|\right)$, i.e., the same as of running (1).

5 EXPERIMENTAL EVALUATION

5.1 Datasets

We investigate this phenomenon in two real-world datasets, which contain both users' ratings and sensitive attributes. The first dataset, Movielens-1M (ML-1M) [4], contains 1,000,209 ratings of $|\mathcal{I}| = 3,952$ movies made by $|\mathcal{U}| = 6,040$ users. Gender and age information is included in this dataset, i.e., $\mathcal{A} = \{gender, age\}$. Specifically, the gender is denoted by a binary attribute[1], $\{m, f\}$. The age is specified in seven ranges, $\{< 18, 18-24, 25-34, 35-44, 45-49, 50-55, > 55\}$.

The second dataset that we used is BookCrossing (BC) [17]. It has 53,408 users, 263,956 items, and 745,161 ratings. This dataset has the attributes age and location provided for each user, $\mathcal{A} = \{age, location\}$. Given the *location* of each user, represented as a tuple containing $(city, region, country)$, we created the demographic groups based on their continent of provenience. This assumption allowed us to obtain groups large enough to assess statistically valid results. To do so, we mapped countries and their respective continent by means of a country-continent table[2]. However, this mapping is not always possible because of the location data provided by users is incomplete. This process led to 22,625 users with valid continent locations. The following continents were identified: **AF** - Africa, **AS** - Asia, **NA** - North America, **SA** - South America, **OC** - Oceania, and **EU** - Europe. We used {EU, AS+OC, NA+SA, AF}. We grouped ages as $\{< 20, 20-40, 40-60, > 60\}$.

[1] While gender is by no means a binary construct, to the best of our knowledge, no dataset with non-binary genders exists.

[2] https://datahub.io/JohnSnowLabs/country-and-continent-codes-list

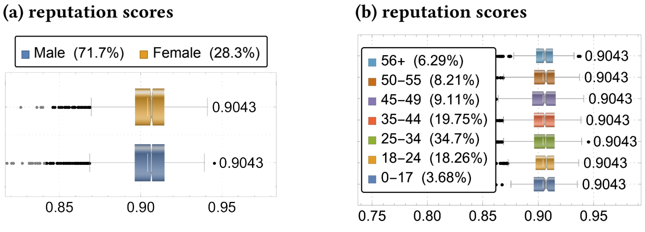

Fig. 3: [ML-1M] Box-whisker-chart for reputations of users resulting from (1) **plus** (3) **in (a) and (b), respectively, with** $\lambda = 0.5$**, for the multiple attributes** **gender** **and** **age.**

5.2 Disparate Reputation Evaluation

Remark 2. *Our approach, in Section 4, works with* any *set of sensitive attributes and the cardinality of* $\mathcal{A}$ *can be higher than 2.* ◇

[ML-1M]. We first characterize the DR, after applying the original method in (1). The results are reported in a Box-whisker-chart (BWC) representing the average reputation for each group, considering gender and age as attributes. Fig. 1 (a) shows us that using solely (1) leads to a consistent reputation disparity on the gender-based groups. On average, male users have higher reputation values than female users. We test the null hypothesis, H_0, for both attributes that the median difference is 0 at the 5% level based on the MW test[3], which is rejected, confirming a gender bias.

Fig. 2 (a) uncovers a disparity on the attribute age, when applying (1). Users belonging to younger groups have, on average, a lower reputation than older users. The DR metric, when only (1) is used, reveal a prominent bias. The MW test for users' reputations, after mitigating for both attributes, return a DR of approximately 0 and, for MW between all pairs, we assess the null hypothesis that the median difference is 0 (H_0) at a 5% confidence level.

When we mitigate bias for both gender and age with (3), we obtain the BWC for reputations under the gender attribute of Fig. 3 (a). We get a DR of $\Delta(a, a') \approx 0$, for each pair of gender-based groups, thus mitigating the bias on the attribute gender. This time, H_0 not rejected. This result confirms that we mitigated the bias on the reputations for these two classes. For the attribute age, we achieve the results in Fig. 3 (b). Now, H_0 is not rejected.

[BC]. First, we assess the disparity originated by the method in (1). In a BWC that considers age and location, Fig. 4 (a) shows that using solely (1) leads to a consistent DR for age-based groups. The younger are the users, the larger average reputation values the class has, thus yielding a bias on the attribute age. To assess this disparity, we test H_0, under an MW test. The results show that the disparity occurs only when the age gap between the users is large and only affects the (less represented) groups of elder users.

Fig. 4 (d) shows the impact of (1) in a BWC for the location attribute. Results surprisingly indicate that, on average, the smallest group (AF) obtains the highest reputation values. We conjecture that this might be because the group might represent a small and cohesive community. The DR values are not zero, showing the presence of bias. After mitigation, the median difference between two classes of the attribute is 0, at the 5% confidence level, under the MW test; a DR only occurs between European and American users, with the latter having a higher average reputation.

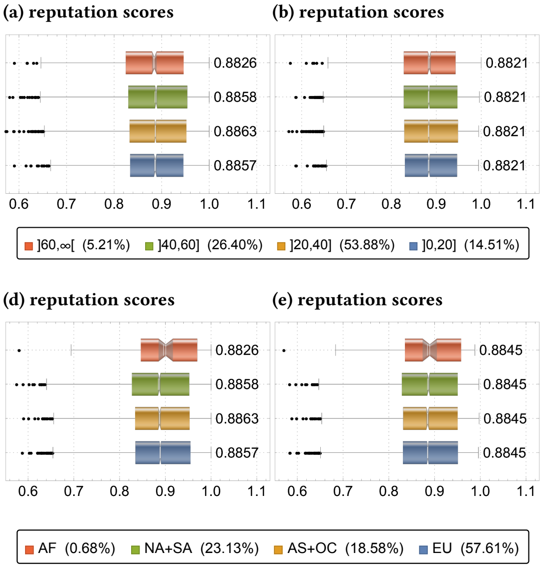

Fig. 4: [BC] Box-whisker-chart for reputations of users after (1) **in (a) and (c), respectively, and after** (1) **plus** (3) **in (b) and (d), respectively, with** $\lambda = 0.5$**, for the multiple attributes** **age** **and** **location.**

Remark 3. *Creating classes with a finer granularity favors the emergence of disparities (see the difference between the age attribute in the two datasets). Besides, multi-class attributes where two of the classes represent the majority of the user base (e.g., location in BC) behave as binary attributes, leading to possibly uncovering disparities only between the two biggest classes.* ◇

When introducing our multi-attribute RI with (3), Fig. 4 (b) shows a DR $\approx$ 0. The MW tests confirm that we cannot reject the null hypothesis and that we can mitigate DR for attribute age. The same occurs for attribute location, as observed in the BWC in Fig. 4 (c).

6 CONCLUSIONS

Automatic computation of user reputations for ranking purposes suffers from biases related to users' sensitive attributes. While recent work has introduced interventions to introduce reputation independence from sensitive attributes, in this work we show that this does not protect users if we group them on a different sensitive attributes. To provide equity guarantees in the reputation scores' computation, we propose an approach to introduce reputation independence from any sensitive attribute of the users. Results on real-world data show that our approach can reach the desired goal.

Future work will lead us to the analysis of the impact on robustness of our solution, the design of group-based metrics to assess the effectiveness of ranking systems, and the inclusion of other benchmark methods and experimental datasets.

Reproducibility. The code is available at https://fenix.tecnico.ulisboa.pt/downloadFile/1407993358910329/Fair_Reputation_Based_Ranking.nb.

Acknowledgments. This work was supported in part by project RELIABLE (PTDC/EEI-AUT/3522/2020) funded by FCT/MCTES.

[3] The DR metric is based on averages obtained from two populations, while the MW is a statistical test that compares the median of the two populations.

REFERENCES

[1] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *International ACM Conference on Research & Development in Information Retrieval*. ACM, 405–414.

[2] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. *CoRR* abs/2004.13157 (2020).

[3] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*. ACM, 2125–2126.

[4] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *ACM Tran. on Interactive Intelligent Systems (TIIS)* 5, 4 (2015), 1–19.

[5] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017 (LIPIcs)*, Vol. 67. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23.

[6] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2019. Search bias quantification: investigating political bias in social media and web search. *Inf. Retr. Journal* 22, 1-2 (2019), 188–227.

[7] Rong-Hua Li, Jeffery Xu Yu, Xin Huang, and Hong Cheng. 2012. Robust reputation-based ranking on bipartite rating networks. In *Proceedings of the 2012 SIAM international conference on data mining*. SIAM, 612–623.

[8] Guilherme Ramos and Ludovico Boratto. 2020. Reputation (In)dependence in Ranking Systems: Demographics Influence Over Output Disparities. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020*. ACM, 2061–2064.

[9] Guilherme Ramos, Ludovico Boratto, and Carlos Caleiro. 2020. On the negative impact of social influence in recommender systems: A study of bribery in collaborative hybrid algorithms. *Inf. Proc. & Management* 57, 2 (2020), 102058.

[10] Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. *CoRR* abs/1901.10437 (2019).

[11] João Saúde, Guilherme Ramos, Ludovico Boratto, and Carlos Caleiro. 2021. A Robust Reputation-Based Group Ranking System and Its Resistance to Bribery. *ACM Trans. Knowl. Discov. Data* 16, 2, Article 26 (July 2021), 35 pages. https://doi.org/10.1145/3462210

[12] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*. ACM, 2219–2228.

[13] Elaine Walster, Ellen Berscheid, and G William Walster. 1973. New directions in equity research. *Journal of personality and social psychology* 25, 2 (1973), 151.

[14] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2019. Fair Learning-to-Rank from Implicit Feedback. *CoRR* abs/1911.08054 (2019).

[15] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *ACM Conference on Information and Knowledge Management*. ACM, 1569–1578.

[16] Meike Zehlike and Carlos Castillo. 2020. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *WWW '20: The Web Conference 2020*. ACM / IW3C2, 2849–2855.

[17] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*. 22–32.